Developing a Teacher Evaluation Instrument to Provide Formative Feedback Using Student Ratings of Teaching Acts

van der Lans, Rikkert M.; van de Grift, Wim J.C.M.; van Veen, Klaas

Published in:
Educational Measurement: Issues and Practice

DOI:
10.1111/emip.12078

Publication date:
2015

Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts

Rikkert M. van der Lans

Wim J.C.M. van de Grift

Klaas van Veen

Department of Teacher Education, Faculty of Social and Behavioral Sciences, University of Groningen, The Netherlands

Correspondence should be addressed to Rikkert van der Lans, Department of Teacher Education, University of Groningen, PO Box 800, 9700AV Groningen, The Netherlands. Telephone: +31 50 363 9754, E-mail: r.m.van.der.lans@rug.nl

Development of a teacher evaluation instrument

Abstract

This study reports on the development of a teacher evaluation instrument, based on students' observations, that exhibits cumulative ordering in terms of the complexity of teaching acts. The study integrates theory on teacher development with theory on teacher effectiveness and applies a cross-validation procedure to verify whether teaching acts have a cumulative order. The resulting teacher evaluation instrument comprises 32 teaching acts with cumulative ordering in terms of complexity. This ordering aligns with prior teacher development research. It also represents a valuable extension, in that the instrument can provide feedback about a teacher's current phase of development and advice for improvement.

Keywords: teacher evaluation, teaching quality, teacher development, teacher effectiveness, Rasch model

Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts

Many Western countries seek to improve education by adopting revised teacher evaluation policies. The drivers of this shift are value-added teacher evaluations (e.g., Firestone, 2014), which are designed to describe the extent to which a teacher has contributed to student achievement gains in a school year. However, value-added evaluations can only inform teachers about their gains; they shed light on neither why students obtained that gain nor how they could improve their gains (Darling-Hammond, Amrein-Beardsley, Heartel, & Rothstein, 2012; Firestone, 2014; Hill, Kapitula, & Umland, 2011). Therefore, current consensus holds that other evaluation instruments are required to complement value-added teacher evaluations.

This consensus has shifted research attention toward further development of classroom observation instruments and student survey instruments as means for evaluation (e.g., Danielson, 2013; Hill et al. 2011; Bill & Melinda Gates Foundation, 2012). Although these instruments are effective in providing more precise information about what a teacher does inside the classroom, such information does not automatically translate into formative feedback (i.e., information about how to develop and improve further). The provision of feedback would require connecting teacher evaluation instruments with teacher development theories. Available teacher development research indicates that the process of becoming an expert teacher follows specific, sequentially or cumulatively ordered phases (Berliner, 2001; Fuller, 1969). Despite widespread acceptance of these theories, the field lacks an evaluation instrument that can provide feedback about which phase of development a teacher has reached and which teaching skills should be considered next for ongoing teacher training, reflection, and self-study. We propose a teacher evaluation instrument that exhibits cumulative ordering and that can provide formative feedback to teachers about their current phase in development.

Theoretical background

The theoretical background is structured in three parts: We consider and summarize teacher development theories; then relate them to key findings about teacher effectiveness; finally, we consider the pros and cons of two evaluation methods; student ratings and classroom observations.

Theories of teacher development

Theories of teacher development describe progressive changes in teacher concerns (Fuller, 1969) as well as a progressive development from novice to expert (Berliner, 2001). From such works we seek to define an ordering that parallels and can be integrated with findings from teacher effectiveness literature. However, we acknowledge though that theories of teacher development traditionally focus on (sequential stages in) teacher *cognition*, rather than teacher *behavior*, which is the focus in teacher effectiveness research. Therefore, our exploration relies on the presumption that teachers' (cognitive) concerns partially reflect observable difficulties and changes they encounter in their teaching. In addition, we note that theory on teacher development, and in particular Fuller's theory, have stimulated two different strands of research (Conway & Clark, 2003); one dedicated to the description of the developmental dynamics of teaching, and one dedicated to the evaluation of teacher concerns in the context of innovation and reform. This paper contributes and is connected with the former; description and measurement of the development of teaching.

Fuller's (1969) theory of teacher concerns is among the first to describe teacher development. It features a relatively simple, three-stage model in which teachers first are concerned with the self, before they turn their attention to tasks, and finally toward students and the impact of their teaching (Conway & Clark, 2003). Concerns for the self center on issues of authority, respect, status, and relationships. Concerns with tasks involve classroom management and content adequacy. Concerns with the impact of student learning pertain to

teachers' ability to specify objectives for students, understand students' capacities, and identify their own contributions to students' difficulties (Fuller, 1969).

Teacher development in Fuller's first two stages, in particular, is well documented. Berliner (2001) describes teachers' growth from novice to expert. For novice teachers, Berliner highlights the importance of developing classroom routines for management and instruction (i.e., tasks). The life-cycle teacher career model (Steffy & Wolfe, 2001) describes six phases, ranging from novice to emeritus, and predicts that teachers who have successfully completed their teacher education begin by developing routines for lesson preparation and achieving reciprocal respect (i.e., task and self). Schafer, Stringfield, and Wolfe (1992) conclude, on the basis of a two-year longitudinal study, that classroom management and basic instruction are among the first teaching skills acquired by teachers (i.e., task).

Regarding the third stage, the impact of student learning, current understanding about its development is limited. The few works exploring the development of more experienced teachers conclude that, in contrast with the relatively homogeneous development of more elementary stages, acquiring skill in the more complex stages is much more varied among teachers, and some teachers never acquire them (e.g., Berliner, 2001; Huberman, 1993).

Discussion also has focused on the rigidity of the proposed stages. Fuller's theory has been characterized as "Perhaps the most classic of stage theories in that it was meant to be relatively invariant, sequential and hierarchical" (Richardson & Placier, 2001, p. 910). In contrast, Berliner (2001), Steffy and Wolfe (2001), and Huberman (1993) suggest a more tentative heuristic interpretation in terms of phases in teacher development. In their view, teachers can develop competence at any time during any phase, and yet at any moment also be grouped into one best-fitting phase. This tentative heuristic approach has the advantage of being less restrictive when describing individual differences in the development of teaching skill, but at a cost: Because it does not exclude any developmental trajectory, information

about current teaching does not reveal the most logical steps for further development and improvement. As Richardson and Placier (2001, p. 913) conclude, "the use of a very flexible approach to stages or phases may have taken us so far from the original concept of a stage theory that the usefulness of the work must be rethought." In contrast, Fuller's invariant, hierarchical approach restricts the individual variation in development of teaching skill, but—if valid—it has the potential to inform an individual teacher about logical steps for ongoing training, reflection, and self-study.

Teacher effectiveness literature and development in teaching skill

Several reviews and meta-analyses address the relation between teaching acts and student achievement (Hattie, 2009; Kyriakides, 2013; Marzano, 2003), and though they use different labels, they show consistently that similar categories of teaching acts enhance student achievement. We consider six broad domains of teaching acts that can be observed within classrooms: creating a safe learning climate, efficient classroom management, quality of instruction, student activation, teaching learning strategies, and differentiation (for in-depth descriptions, see Appendix A). Author (2013) provides an extensive literature review to account for the six domains. In addition, Authors, et al. (2014) describe connections between these six domains and the classroom assessment scoring system (CLASS) and the framework for teaching (FFT) observation protocol, both of which are currently employed in the Measures of Effective Teaching (MET) project. They conclude that the six domains coincide with all the clusters of the FFT and CLASS.

Appendix A compares the six domains with the seven Cs of the Tripod survey (Bill & Melinda Gates Foundation, 2012), a student questionnaire employed in the MET project. The survey is clustered into seven factors—caring, controlling, clarifying, challenging, captivating, conferring, and consolidating—that measure how students experience the teacher's behavior. As Appendix A shows, descriptions of the seven Cs coincide with the

descriptions of the six domains. The overall impression is that the seven Cs coincide with four domains: safe learning climate, efficient classroom management, quality of instruction, and activating students. The learning strategies and differentiation domains appear relatively unique to our framework.

In addition, Appendix A notes possible connections between the six domains and Fuller's three stages of teachers' concerns. We acknowledge that these connections are to some extent speculative, but they may contribute to an understanding of the six domains in terms of progressive stages. Our speculations are based on some recent empirical studies (Kyriakides, Creemers, & Antaniou, 2008; Authors, et al. 2011) that indicate a cumulative ordering of teacher behaviors, from less to more complex, which may reflect teaching development. Kyriakides, Creemers, and Antaniou (2008) group teaching acts into five types and find a cumulative ordering that gradually moves from actions associated with direct teaching to more advanced actions involving new teaching approaches and differentiation. Authors, et al. (2011) analyze classroom observations performed by trained colleagues in elementary education of the identical six domains of teaching acts. This study found they are cumulatively ordered, from a safe learning climate to efficient classroom management to quality of instruction to student activation and finally to differentiation and then learning strategies.

Evaluation method: student ratings

The success of an evaluation instrument depends on its ability to present feedback to individual teachers about their teaching. This criterion creates some different and unusual demands. Unlike the conventional goal of empirical research—to generalize across people—our focus is on generalizing across situations in which a person acts. Furthermore, the chosen method ideally has low implementation costs but still provides feedback that is informative about a relatively wide range of situations. With these considerations, we discuss the

advantages and disadvantages of two observational methods: classroom observations and student ratings.

Classroom observations. A classroom observer may be a trained assessor or someone with extensive experience observing classrooms. The principal advantage of classroom observation is that the observer is not involved in any way in the lessons. Ideally, well-trained observers evaluate teachers using a similar norm and therefore should be more objective (Muijs, 2006). However, a single observation cannot reflect the teacher's average performance over a larger set of situations. To achieve reliable estimations of performance across time, some studies recommend three to six classroom observations (e.g., De Jong & Westerhof, 2001; Hill, Charalambous, & Kraft, 2012). Another disadvantage of this method is the potential for observer bias. If only one observer evaluates the teacher on multiple occasions, those observations could reflect the observer's prejudices and personal values; interaction effects between observers and teachers also could clutter the evaluation results. The solution would be to have multiple observers assess the teacher (Peterson, 2000). Overall then, classroom observation offers the advantages of an objective, outside perspective, but it requires the use of multiple trained observers who observe each teacher on three to six occasions in each class. For schools to adopt classroom observations for their teacher evaluations, the costs would likely be enormous, while the benefits yet remain uncertain.

Student ratings. Researchers and teachers have long been suspicious of student ratings. Because students are closely involved in the lessons, they are not independent or objective raters. However, most recent research indicates that student ratings can provide trustworthy, valid insights for teacher evaluation (Marsh, 2007). An advantage of student ratings is that they usually span many observers at once, thereby substantially decreasing observer bias (Marsh, 2007; Richardson, 2005). In addition, research shows that students ratings vary primarily as a function of the teacher's teaching skill (Benton & Cashin, 2012; Richardson,

2005). Furthermore, student ratings tend to be stable over time (Benton & Cashin, 2012), which suggests that students rate teaching acts according to their average perception across all previous encounters. These advantages make student ratings considerably more cost effective than classroom observations. Concerns with student ratings mostly involve the potential for bias. Researchers have directed considerable attention to bias due to students' expectations about their grades (i.e., whether students favor lenient graders) and due to students' prior interest in the subject matter (i.e., whether students misattribute their own subject matter interest to be caused by the teacher), but these biases are generally small (Benton & Cashin, 2012; Marsh, 2007; Richardson, 2005). More profound concerns relate to student expertise; younger students in particular may not be aware of valuable information required to evaluate teachers (Peterson, 2000). Furthermore, students are not trained observers, and compared with classroom observers, they have relatively little experience with differences in teaching. In summary, student ratings offer a relatively cost-effective evaluation method, because they are unpaid evaluators and require few evaluation moments, but the evaluations reflect what students expect from the teacher, not a trained preset, standardized norm.

Against this background, we address the following research questions: *Can we establish a cumulative order of effective teaching acts, using student observations?* And; *How may the development of such a scale contribute to the knowledge about teacher development?*

Method

Sample

The sample for this study consisted of 2,262 student ratings, obtained from a school for secondary education in the Netherlands (student ages: 12–18 years). Female students constituted 53.1% of the student sample (1,200). The school offers vocational, higher vocational, and pre-university education. Students judged 68 teachers working at the school.

The study included teachers from all subjects except Physical Education. Teaching experience ranged from 0 to 43 years, with an average of 16 years.

Measurement instrument

The survey includes items reflecting 59 teaching acts, such as "This teacher knows what I'm able to do" or "This teacher ascertains that I understand the subject matter taught." Students rated each teaching act on a dichotomous response format (1 = "rarely observed" to 2 = "often observed"). We deliberately chose for dichotomous response format, because in the Rasch model its interpretation is more straightforward and though the feedback is more easily explained to teachers. Simplicity is perceived key to implementation.

Design and missing values

In the nested design, the aggregate level identifies 84 unique teacher–class combinations. This number exceeds the number of teachers in the data set because for some teachers, ratings were available from two classes, resulting in two unique combinations. Of the 131,458 item responses given, 2,016 were reported missing, a 1.5% rate of missing values. We considered these missing values to be missing at random (MAR).

Cross-validation procedure

The method relied on a cross-validation procedure for which the complete sample was split into development and validation samples. The complete sample counted 2,262 student ratings. We established a development sample by randomly selecting 10 students from each teacher–class combination ($n_{development}$ = 840). This development sample served to calibrate the measurement instrument.

To establish the validation sample, we randomly selected another 10 students from each teacher–class combination ($n_{validation}$ = 750). The validation sample was slightly smaller than the development sample because a few classes contained fewer than 20 students, so

fewer than 10 students remained for the validation sample; six teacher–class combinations had fewer than 6 student ratings left to include in the validation sample. To limit sample imbalance, we excluded these combinations, such that the validation sample also featured six fewer teachers than the development sample.

In total, 1,590 students are included in the development and validation samples. The other 672 students were omitted. Subsamples did not differ in student total test scores ($F(2, 2214.58) = .27$, $p = .79$) or in student age ($F(1, 2193.89) = .20$, $p = .66$), though they did differ slightly on student gender ($\chi^2(2, N = 2{,}226) = 6.90$, $p = .03$). The omitted sample had 57.8% girls, while the two randomly selected samples; 51.3% and 52.4%.

Model specification

Our research question pertains to whether we can find a cumulative order for effective teaching acts. To address it, we apply the Rasch model, generally considered the most appropriate model to test for cumulative item ordering (Bond & Fox, 2007). The Rasch model relies on three assumptions (DeMars, 2010):

1. *Parallel item characteristic curves (ICCs).* This assumption states that each teaching act can discriminate equally among levels of teaching skill.
2. *Unidimensionality.* This assumption states that student responses can be ascribed to a single latent construct: teaching skill.
3. *Local independence.* This assumption states that the residuals of item pairs are uncorrelated.

We deliberately chose the strict one-parameter item response theory (IRT) model (i.e., the Rasch model) instead of the two-parameter IRT model. We view the two-parameter IRT model as an effective option to develop latent measurement scales, but it cannot be applied to test for cumulative ordering, as is examined in this study (Bond & Fox, 2007).

The Rasch model can be understood as a generalized linear mixed model specifying two components (De Boeck, et al. 2011):

$$log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \sum_{j=1}^{J}\theta_j + \sum_{i=1}^{I} b_i$$

We refer to the first of these components as the "structural model" and the second as the "measurement model." Interactions between the components suggest model violations. The plus sign signals that b_i should be interpreted as item easiness. If—as in our case—the design is nested and a third component is added to this equation, we must specify whether the third component is nested within the structural model or within the measurement model. In this study, we view students as nested in teachers, and they together define the structural model. Because items are not nested, their fit is assessed by application of the regular single-level item fit statistics. As we discuss at the end, we view this approach as defendable yet not entirely satisfactory.

Data analysis

The analyses consist of two sections: (1) validation of the measurement model and (2) examination of the structural model. The validation of the measurement model is further subdivided in two subsections: development and validation phases.

Validation of the measurement model. In the development phase, we tested for item fit with the three Rasch model assumptions. We excluded from further analysis any item that did not meet any one of these assumptions. Test included are; Andersen (1973) likelihood ratio (LR) test to evaluate the assumption of parallel ICC, exploratory factor analysis (EFA) to evaluate the assumption of unidimensionality, and Ponocny's (2001) nonparametric T_1 and T_{1m} to evaluate the assumption of local independence.

In the validation phase, we reassessed the fit of the remaining items to ensure that the teaching acts had not been selected on the basis of chance. This second phase is directed at

validation, not item selection. The validation involved identical tests with exception of the EFA; we consider confirmative factor analysis (CFA) more appropriate for validation.

Structural model: An exploration of measurement reliability. In this section, the results involve the measurement reliability and marginal standard error of measurement (SEM). Following Raju, Price, Oshima, and Nering (2006), we estimate the group reliability for teachers ($\rho_{(\theta\theta')T}$) and students ($\rho_{(\theta\theta')S}$). Analogous to Patz, Jucker, Johnson, and Mariano (2002), we turn to the hierarchical structure and explore how raw scores are translated into different values of θ scale and its associated SEM. However, unlike Patz et al. (2002) but consistent with Brennan (2004), we do not interpret the rater facet as constituting bias or rater severity. Variation in students' ratings is equally interesting and may ultimately prove useful in informing teachers about possible steps to improve their teaching with regard to particular target students; however, the scope of this discussion transcends the primary goal of this article: to develop a Rasch-scaled student rating instrument for teacher evaluation.

Software

The data analysis procedure relied on R and Mplus version 7 (Muthen & Muthen, 1998–2012). In R we installed the eRm R-package (Mair & Hatzinger, 2007), which uses a conditional maximum likelihood algorithm to estimate the item fit statistics. Mplus applies a robust weighted least squares estimator algorithm to estimate item fit. The nested components of the structural model were estimated with the R package lme4 version: 1.1-7 (Bates Maechler, Bolker, & Walker, 2014).

Results

We begin this section by presenting the results for measurement model. Starting with the instrument calibration in the development sample, then a reexamination of item fit in the validation sample and ending with a presentation of our proposed evaluation instrument. The result section then turns to the structural model and explores measurement reliability.

Development sample

Parallel ICC

Anderson (1973) proposes an LR test of parallel ICCs, splitting observed data into two subgroups: one that scores low on the measured latent trait (i.e., low teaching skill) and another that scores high on it (i.e., high teaching skill). The LR test then compares the deviance in the log-likelihood ratios of both groups against a chi-square distribution. We performed the median as the split criterion. The LR test revealed that not all 59 teaching acts achieved parallel ICC ($\chi^2 = 286.10$, df = 58, $p = .00$). Therefore, we excluded the teaching act that resulted in the greatest decrease in the chi-square value over repeated rounds, until 43 of the initial 59 teaching acts remained; on average, they exhibited parallel ICC ($\chi^2 = 54.50$, df = 42, $p = .09$).

Unidimensionality

The unidimensionality assumption is difficult to (dis)confirm (DeMars, 2010). All measurement instruments are, to some extent, multidimensional, and we can only test whether unidimensionality is defensible. A common strategy uses factor analysis, which suggests that, provided unidimensionality holds, the best factor solution of the correlations among the 43 teaching acts should be a one-factor solution. We used tetrachoric correlations, because Pearson phi correlation coefficients can prompt high loadings for ratings with similar difficulty (DeMars, 2010). The eigenvalues of the EFA, as plotted in Figure 1, suggest a one-factor solution. The first eigenvalue (21.23) is considerably larger than the second (2.01) and third (1.89) eigenvalues.

INSERT FIGURE 1 AROUND HERE

Local independence

To test the local independence assumption, we used Ponocny's (2001) T_1 and T_{1m}. Rasch (1960) was especially concerned about this third assumption of his model and

originally proposed, but never completed, a nonparametric test to assess model fit. Ponocny's (2001) family of T-statistics implements some of Rasch's original design. The T-statistics specify each an one-tailed directional alternative hypothesis, which increases their power considerably. The T_1 statistic evaluates violations of local independence due to increasing (i.e., positive) residual correlations, and the T_{1m} statistic evaluates violations due to decreasing (i.e., negative) residual correlations.

Chance should have an important position in evaluating the T-statistics results (I. Ponocny, personal correspondence, September 30, 2014). The T-statistics pair every item with 42 other items. Therefore, a criterion of two violations per item would reflect an alpha criterion of .05. However, their considerable power together with the slight overlap in item content (both within and between domains) and students' differential grammar ability, makes that some additional violations are almost inescapable and may be tolerated. On the basis of these considerations we decided to set a more lenient criterion of 5 violations.

Acts 5 ("My teacher explains well") and 51 ("This teacher makes sure I understand his/her explanation") together yielded 33 of the total 109 violations for T_1. Moreover, act 15 ("My teacher asks questions that make me think") alone accounts for 25 violations for T_{1m}. Continuing with the calibration, we deleted additional teaching acts over repeated rounds, starting with the act that accounted for the most violations. After excluding 13 teaching acts, the 32 remaining acts combined for 37 violations due to increasing correlations and 28 violations due to decreasing correlations and no act accounted for more than 5 violations.

Validation sample

We reexamined the fit of the 32 teaching acts with each of the Rasch model assumptions using the validation sample ($n_{validation}$ = 750). The Andersen LR test confirmed that, on average, all teaching acts achieved parallel ICC (χ^2 = 36.90, df = 31, p = .22). A CFA, applied to reexamine the unidimensionality assumption, showed that the one-factor model fit

the data well (root mean square error of approximation = .029, confirmatory fit index = .96, Tucker–Lewis index = .96). The scree plot confirmed that the one-factor solution was defensible. Finally, with regard to local independence, Ponocny's (2001) T_1 indicated that four teaching acts had more than 5 violations, and T_{1m} indicated that three teaching acts had more than 5 violations (see Table 1). In total, 7 of the 32 teaching acts failed to meet the local independence criterion in the validation analysis, but these 7 violations did not seem to cluster around any particular domain. Overall, we consider these results encouraging.

In addition, we examined the invariance of the item (*b*) parameters between the development and validation samples. Figure 2 shows the goodness-of-fit plot. The 32 dots indicate the 32 items, the dashed line reflects the perfect invariance between samples (i.e., the zero-difference score). The deviations from the dashed line indicate deviations from item invariance. The goodness-of-fit plot shows that—with the exception of act 12 ("My teacher treats me with respect")—the item parameters can be considered invariant between samples.

INSERT FIGURE 2 AROUND HERE

Final questionnaire

We present the established scale in Table 1. The *b* coefficients indicate the difficulty (i.e., here complexity) of the teaching act, such that low values signify teaching acts with less complexity. Because these 32 teaching acts fit our criteria for cumulative ordering, it follows that the acts with higher *b* coefficients could have been rated "often" by students only if (most) teaching acts with lower *b* coefficients also were rated "often". Thus, the less complex teaching acts can be considered prerequisites for more complex teaching acts.

INSERT TABLE 1 AROUND HERE

Broadly, the cumulative ordering in Table 1 aligns with descriptions of teacher development: It starts with teaching acts that establish a safe learning climate and quality of instruction and ends with teaching acts associated with differentiation and learning strategies.

This result confirms our predicted cumulative ordering in complexity. Furthermore, the ordering in Table 1 shows considerable within-domain variation, for efficient classroom management in particular. This result suggests that the least complex skills of (more complex) domains may precede the development of the most complex skills of other (less complex) domains. This finding fits with discussions about the limitations of perceiving teacher development in rigid stages, which have continually suggested that descriptions (and measurement) of development in invariant stages is inappropriate and that more flexibility is desirable. By establishing the cumulative ordering at the level of acts, the instrument avoids the requirement of a complete invariant hierarchical ordering in domains. We also note that the ordering includes teaching acts from all six previously identified domains of effective teaching. Our strict procedures for selecting teaching acts thus did not exclude any domain from the instrument; omitting 27 teaching acts seemingly did not produce any unacceptable loss of information. In support of this assertion, we computed the correlations of the evaluation scores for teaching skill measured with the original 59 teaching acts versus those measured by the 32 selected teaching acts. A high correlation would suggest that excluding the 27 teaching acts had a minor impact on final evaluations of teaching skill. Indeed, we find that the Pearson product moment correlation between teacher skill scores obtained from the 59- versus 26-teaching act instrument was $r = .99$, with $n = 84$ and $p < .00$.

Measurement reliability

To further explore the instrument's properties, we estimated the group-level reliability and SEMs. The group-level reliability (Raju et al. 2006) has similar interpretation to Cronbach's alpha; for students, $\rho_{(\theta\theta')S} = .80$, and for teachers, $\rho_{(\theta\theta')T} = .86$. This result suggests that the instrument reliably discriminates between teachers of different skill. Table 2 presents the local SEM estimates associated with the 32 possible response vectors (response vectors with missing values were omitted). The results suggest increasing measurement precision for

skill estimates located more near the center of the measurement scale. For teachers, measurement precision also depend on the number of raters. The Table 2 further reveals a ceiling effect for the individual student response vectors. It seems thought, that the discrimination between teachers relies on those 71.1% of the students not rating the teacher as "perfect". This is an issue of concern.

INSERT TABLE 2 AROUND HERE

Conclusion and discussion

Our results confirm the main premise: Effective teaching acts can be ordered cumulatively, from basic to more complex. Broadly, the cumulative ordering observed is in accordance with Fuller's (1969) theory on teacher development, which states that teachers are first concerned with the self, then with the task, and finally with their impact on student learning. Thereby, the validation of a cumulative ordering also provides some initial insights in the development of effective teaching behaviors. These findings represent an important step toward instruments that can provide truly formative feedback. In the future, the instrument developed here could provide an alternative to those in use currently, which can score teachers' current skill but lack the underlying, empirically validated, cumulative ordering required to present objective advice about the next steps to improve.

Other potential advantages of the instrument

In addition to the cumulative ordering, our instrument has some advantages over other survey instruments. First, in the future, test for differential item functioning, allow researchers to test for measurement invariance between subgroups. Thereby it can open an empirical discussion about whether all teachers develop and can be evaluated with the same instrument and the same ordering or whether different (yet parallel) instruments for specific subgroups are required (e.g., math teachers, language teachers). Second, the IRT tradition has covered considerable ground in developing person fit tests. These tests may be applied to evaluate

whether any individual teacher deviates from the here presented ordering. Several authors have speculated on individual differences in teachers' developmental trajectories (e.g., Berliner, 2001; Huberman, 1993), and person fit tests may provide useful tools to evaluate and track deviations. Third, our approach offers the possibility for adaptive and tailored testing. If schools adopt the instrument and already have information about the teacher, they do not need to score the entire instrument. Instead, they could choose to focus on the most relevant acts to maximize information with minimal time and effort.

Limitations

We note that the limited sample size of only one school restricts generalization of our findings to other contexts. The results should be viewed in a broader attempt to validate the proposed instrument and its underlying theory. Recently, Authors et al. (2014) published their findings for a sample of student teachers.

Further methodological considerations

We estimated item fit without consideration of the second teacher level. Our rationale is that in IRT models, there should be a strict separation between the measurement and the structural model. In IRT, and specifically in Rasch models, no interaction between model parts is allowed. In multilevel extensions however, this strict separation is more difficult to attain. We plotted a Venn diagram representing the three facets involved and their variances. The dashed circle represents the fixed item effect, and the solid lines represent the random teacher (wider circle) and student (inner circle) effects. As Figure 3 shows, the item × student interaction is negligible and, from an IRT perspective, well handled by the model. However, the item × teacher interaction, though small, is not negligible. This violates the assumed strict separation. We present this result to urge the development of multilevel IRT *item* fit tests, which—to our knowledge—are currently not available in IRT software.

INSERT FIGURE 3 AROUND HERE

Contribution to research

Although the primary goal of this article has been to validate a student survey instrument, the findings also confirm our main premise of stagewise development. This premise has important implications for research. An important one pertains to the interrelations across studied domains and acts. If we accept cumulative ordering, it follows that more complex acts should not be described or measured in isolation from less complex acts, whereas less complex acts may be measured without considering the more complex acts.

Practical relevance

Finally, we aim to contribute to discussions of teacher evaluation by devising an instrument for formative feedback that integrates teacher effectiveness literature with theory on teacher development. We believe this instrument offers great potential to contribute to the provision of formative feedback in teacher evaluations. From an evaluation perspective, the main advantage of this instrument is not simply that it ranks teaching acts by complexity but that, because the ordering is cumulative, the teacher's current position reveals what we might call the teacher's "zone of proximal development."

To graphically portray this result, we plotted the hierarchy in a person–item map. The acts are sorted by domain, from right to left: climate, management, instruction, activation, differentiation, and learning strategies. The less complex domains on the left-hand side are ranked lower on the scale than the more complex domains. Teacher positions (O) are portrayed on the right-hand side of the y-axis. Acts located near the teacher's position are considered most relevant for further training and self-reflection. Acts that exceed the teachers' skill are considered too complex to train; acts below the teachers' skill are already developed. The student positions (X) may eventually appear relevant to present feedback to teachers about how to approach each individual student.

INSERT FIGURE 4 AROUND HERE

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using S4 Classes*. R package version 1.1-7. URL: http://CRAN.R-project.org/package=lme4.

Berliner, D. (2001). Learning about learning from expert teachers. *International Journal of Educational Research, 35,* 463–483.

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: a summary of the research and literature.* (IDEA Paper No. 50). Retrieved March 3, 2015, from http://www.ntid.rit.edu/sites/default/files/academic_affairs/Sumry%20of%20Res%20%2350%20Benton%202012.pdf.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* London: Lawrence Erlbaum Associates.

Brennan, R. L. (2004). *Some perspectives on inconsistencies among measurement models.* (CASMA Research Report No. 10). Retrieved March 9, 2015, from http://www.uiowa.edu/~casma/NSF-casma-rpt.pdf.

Conway, P. F., & Clark, C. M. (2003). The journey inward and outward: a re-examination of Fuller's concerns-based model of teacher development. *Teaching and Teacher Education, 19,* 465–482.

Danielson, C. (2013). *The framework for teaching: Evaluation instrument.* Princeton, NJ: The Danielson Group.

Darling-Hammond, L., Amrein-Beardsley, A., Heartel, E., & Rothstein J. (2012). Evaluation teacher evaluation. *Phi Delta Kappan, 93,* 8–15.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Abe, H., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39,* 1–25.

De Jong, R., & Westerhof, K.J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research 4,* 51–85.

DeMars, C. (2010). *Item response theory. Understanding statistics measurement.* New York, NY: Oxford University Press.

Firestone, W. A. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher 43,* 100-107.

Fuller, F. (1969). Concerns of teachers: a developmental conceptualization. *American Educational Research Journal 6,* 207–226.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Abingdon, Oxon, UK: Routledge.

Hill, H., Charalambous, C. Y., & Kraft, M. A. (2012). When interrater-reliability is not enough: Teacher observation systems and a case for the generalizability theory. *Educational Researcher 41,* 56–64. doi: 10.3102/0013189X12437203.

Hill, H., Kapitula, L., & Umland, K. (2011). A validity argument to evaluating teacher value added scores. *American Educational Research Journal 48,* 797–831.

Huberman, M. (1993). *The lives of teachers.* New York, NY: Teachers College Press.

Kyriakides, L. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education 36*, 143–152.

Kyriakides, L., Creemers, B. P. M., & Antaniou, P. (2008). Teacher behavior and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education 25,* 12–23.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modelling: the eRm package for the application of IRT models in R. *Journal of Statistical Software 20*, 1–20.

Marsh, H. D. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht, The Netherlands: Springer.

Marzano, R.J. (2003). *What works in schools. Translating research into action.* Alexandria, VA: ASCD.

Authors (2014). Blinded for review.

Bill & Melinda Gates Foundation (2012). *Asking students about teaching: Student perception surveys and their implementation.* Retrieved March 3 from http://www.metproject.org/downloads/Asking_Students_Practitioner_Brief.pdf.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation: An International Journal on Theory and Practice 12*, 53–74.

Muthen, L. K., & Muthen, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthen and Muthen.

Patz, R. P., Jucker B. W., Johnson, M. S., & Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341–384. doi: 10.3102/10769986027004341

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practice.* Thousand Oaks, CA: Corwin Press.

Ponocny, I. (2001). Non-parametric goodness-of-fit tests for the Rasch model. *Psychometrika 66,* 437–460.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen & Lydiche.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2006). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement, 30,* 1–12. doi: 10.1177/0146621606291569.

Richardson, J. T. E. (2005).Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education, 30,* 387–415. doi: 10.1080/02602930500099193.

Richardson, V., & Placier, A. (2001). Teacher change. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed.). Washington, DC: American Educational Research Association.

Schafer, E., Stringfield, S., & Wolfe, D. (1992). Two-year effects of a sustained beginning teacher induction program on classroom interactions. *Journal of Teacher Education, 43,* 203–214.

Steffy, B. E., & Wolfe, M. P. (2001). A life-cycle model for career teachers. *Kappa Delta Pi Record 38,* 16–19. doi: 10.1080/00228958.2001.10518508.

Authors (2013). Blinded for review.

Authors, et al. (2011). Blinded for review.

Table 1.

Fuller stage, domain, act, and complexity (b) of 32 teaching acts (n = 1,590)

Stage	**Category**	**Act**	**description**	*b*	SE
Self	Climate	12	Treats me with respect.	–1.32	.176
Task	Management	49	Prepares his/her lesson well. [b]	–.98	.166
Self	Climate	27	Ensures that others treat me with respect.	–.72	.160
Self	Climate	17	Answers my questions.	–.69	.159
Self	Climate	32	Ensures that I treat others with respect.	–.67	.159
Task	Management	50	Makes clear what I need to study for a test.	–.65	.158
Task	Management	58	Helps me if I do not understand or am unable to do something.[a]	–.56	.156
Task	Instruction	57	Uses clear examples.[a]	–.55	.156
Task	Management	31	Ensures that I know what to do.	–.49	.155
Task	Management	6	Ensures that I behave well.	–.36	.153
Task	Management	33	Explains the purpose of the lesson.[a]	–.22	.150
Task	Instruction	14	Explains everything clearly to me.	–.22	.150
Task	Activation	4	Involves me in the lesson.	–.21	.150
Task	Activation	26	Encourages me to think for myself.	–.20	.150
Self	Climate	1	Ensures that I am relaxed in the classroom.	–.14	.149
Impact	Activation	20	Stimulates me to think.	–.10	.148
Impact	Activation	28	Ensures that I pay attention.	–.08	.148
Task	Management	42	Makes clear when I should have finished an assignment.	.00	.147
Impact	Management	43	Applies clear rules.	.04	.147
Task	Instruction	3	Ensures that I know the lesson goals.	.18	.145
Task	Activation	39	Stimulates my thinking.	.25	.144
Impact	Differentiation	35	Connects to what I know or am capable of.	.51	.141
Task	Management	29	Ensures that I keep working.	.53	.141
Task	Management	7	Ensures that I use my time effectively.[a]	.57	.141
Impact	Learning strategies	40	Explains how I should study something.	.63	.140
Impact	Differentiation	16	Checks whether I understood the subject matter. [b]	.72	.140
Impact	Activation	54	Evokes interest	.76	.139
Impact	Differentiation	8	Keeps track of what I know and am capable of.[b]	.78	.139
Impact	Learning strategies	22	Teaches me to check my own solutions.	.81	.139
Impact	Learning strategies	23	Teaches me to simplify problems.	1.06	.137
Impact	Differentiation	47	Knows what I find difficult.	1.36	.135
Impact	Learning strategies	48	Teaches me to summarize what I have read in my own words.	1.68	.134

[a] These items had more than five violations for the local independence assumption due to positive increasing correlations in the validation sample.

[b] These items had more than five violations for the local independence assumption due to negative decreasing correlations in the validation sample.

Table 2.

Marginal estimates of the SEM as a function of θ for students and teachers.

	Students			Teachers			
Raw score	***freq. obs.***	**M(θ)**	**SEM**	***freq. obs.***	$n_{(raters)}$	**M(θ)**	**SEM**
1	0	—	—	0	—	—	—
2	0	—	—	0	—	—	—
3	1	–2.79	.662	0	—	—	—
4	3	–2.66	.572	0	—	—	—
5	3	–2.96	.555	0	—	—	—
6	2	–1.83	.534	0	—	—	—
7	3	–3.22	.527	0	—	—	—
8	2	–2.07	.510	0	—	—	—
9	5	–1.97	.505	0	—	—	—
10	9	–1.81	.504	0	—	—	—
11	13	–2.08	.500	0	—	—	—
12	5	–1.79	.518	0	—	—	—
13	9	–1.68	.516	0	—	—	—
14	10	–1.59	.499	0	—	—	—
15	17	–2.12	.520	0	—	—	—
16	8	–1.76	.519	1	18	–2.34	.347
17	15	–1.50	.511	2	28	–2.12	.407
18	14	–1.68	.509	0	—	—	—
19	26	–1.59	.514	0	—	—	—
20	20	–1.44	.521	1	15	–1.63	.380
21	24	–1.36	.541	1	16	–1.53	.367
22	44	–1.29	.537	1	2	–.86	.737
23	36	–1.26	.536	0	—	–	—
24	44	–.95	.542	6	97	–1.02	.373
25	45	–.91	.561	7	93	–.81	.415
26	69	–.62	.580	10	142	–.59	.408
27	74	–.50	.599	14	246	–.36	.370
28	84	–.33	.624	9	141	–.14	.406
29	112	–.05	.661	9	160	.17	.381
30	103	.17	.733	14	247	.56	.399
31	160	.56	.842	6	93	1.19	.461
32	391	1.26	NA	3	53	1.69	.465

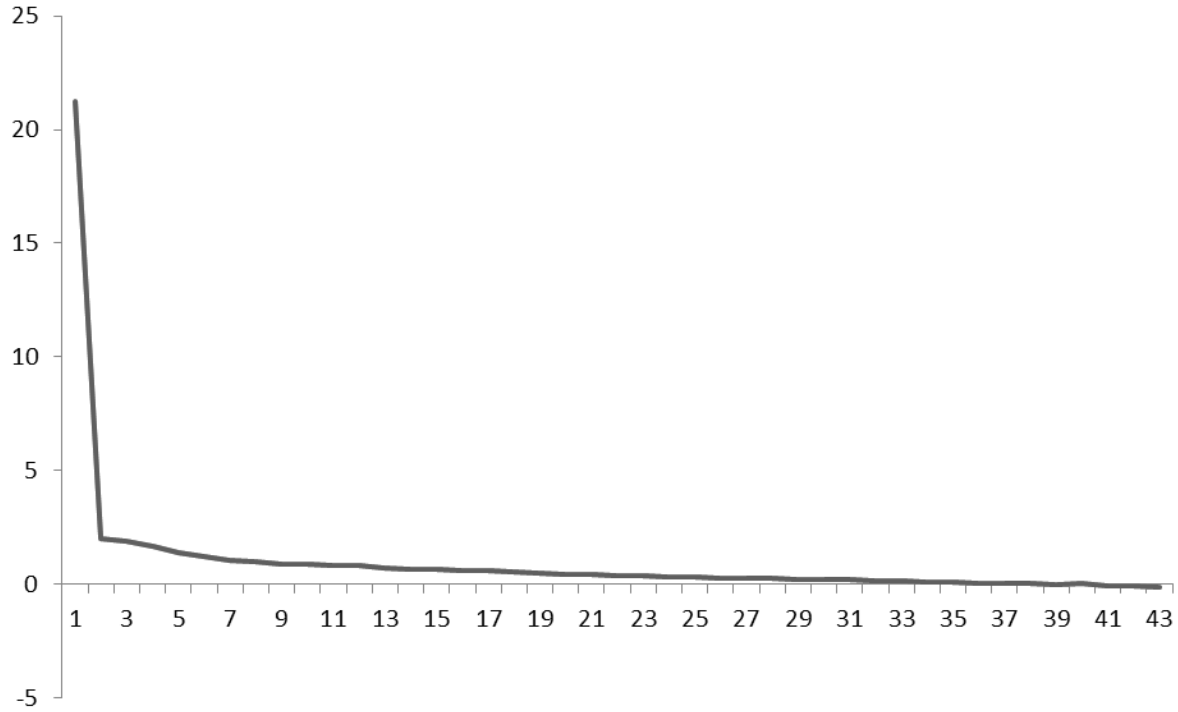

Figure 1. Scree plot of the exploratory factor analysis using the tetrachoric correlations.

Notes: The y-axis shows the eigenvalue, and the x-axis indicates the number of factors.

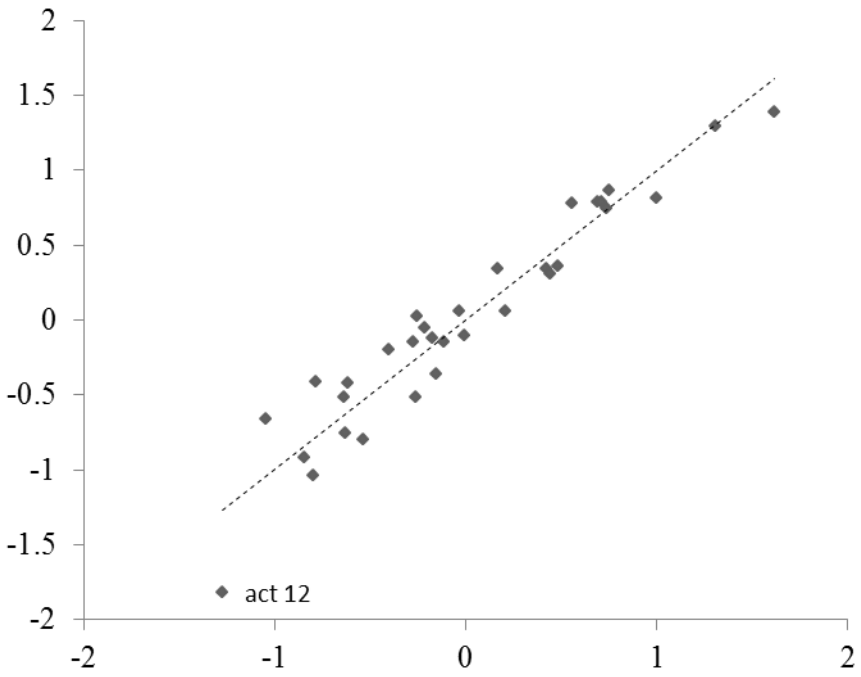

Figure 2. Goodness-of-fit plot visualizing item parameter invariance between the development and validation samples.

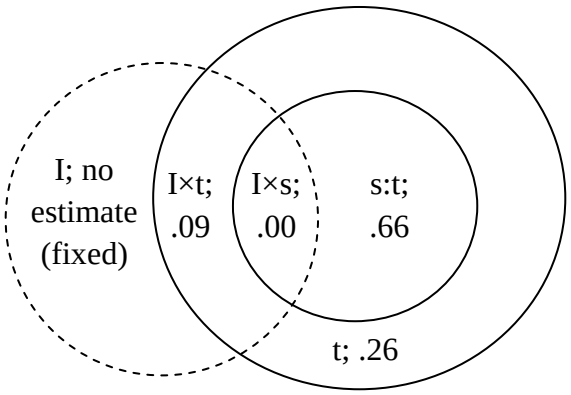

Figure 3. Venn diagram representing the variance decomposition (%) of the facets teacher (t), students nested in teachers (s:t), and item (I) and their interactions.

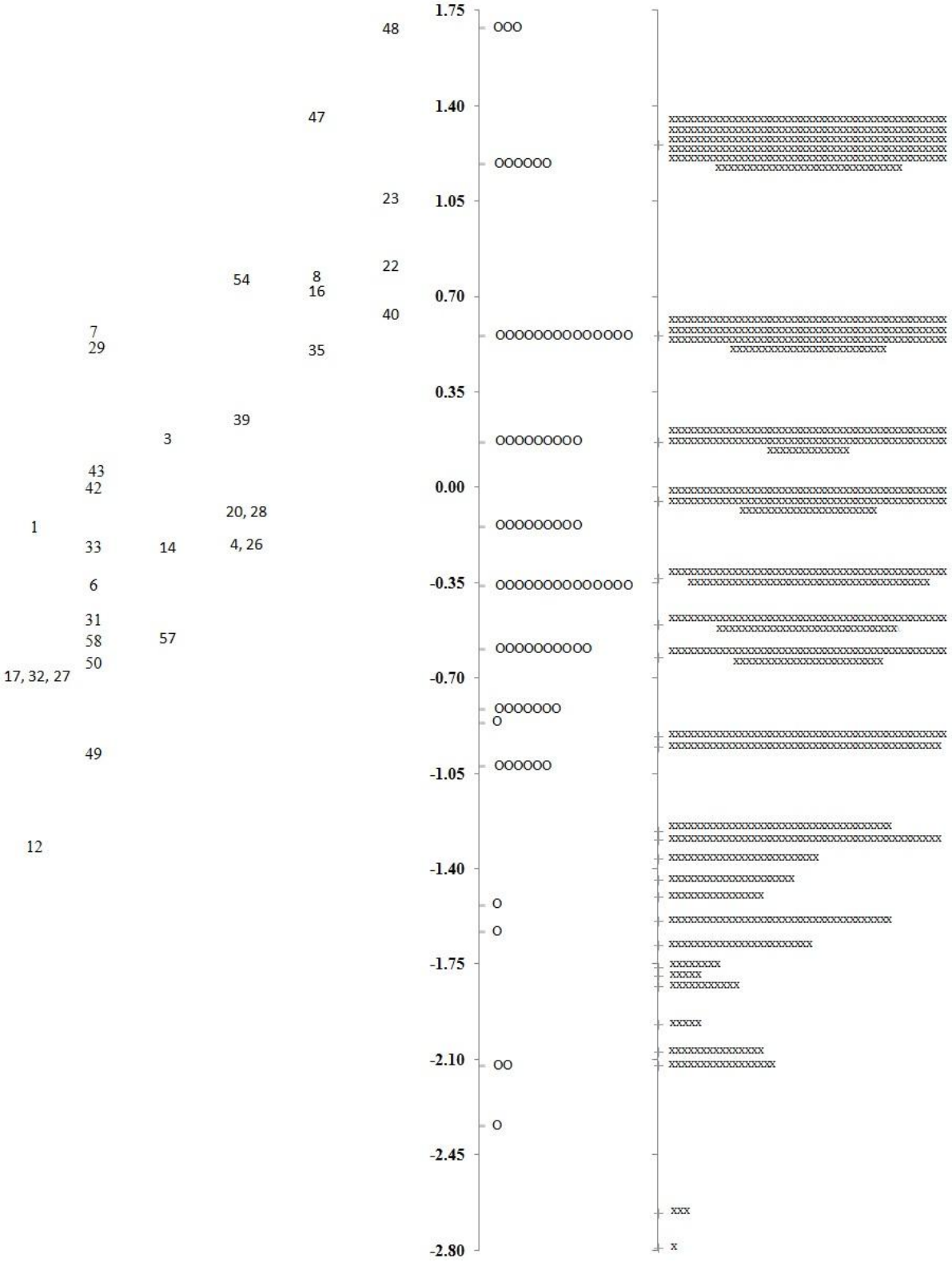

Figure 4. Person–item map. The y-axis gives the range of θ. The right-hand side plots the 32 items' positions. The left-hand side of the y-axis plots the positions of teachers (denoted by "o") and students (denoted by "x") on the measurement scale θ.